

A bivariate analysis for Snedecor's sugar beet example on covariance

BU-52-M

W. T. Federer

May, 1954

Prof. G.W. Snedecor presents an unusual example involving covariance in the discussion on covariance analyses in his "Statistical Methods" book. In covariance analyses interest is centered on the dependent variate, Y, and the independent variate, X, is considered extraneous except as its variation affects the dependent variate. However, it might be that both variates are of interest, that the experimenter wishes to make use of both characteristics in assessing the value of the different treatments, and that one wishes to use a multivariate analysis instead of a covariance analysis (5, 6, 7). For experiments like the Snedecor sugar beet example (4, sec. 12.7) it is quite conceivable that the two characters, stand and yield, are of more importance than yield adjusted for variations in stand. In this connection one should remember that the average weight per individual sugar beet multiplied by the total number of beets per plot represents the total yield per plot. Thus, one could use the two characteristics, average weight per sugar beet per plot and the number of plants per plot, in place of the two characteristics total weight of sugar beets per plot, X_1 , and total number of plants per plot, X_2 . The latter two characteristics are the ones used below for the bivariate analysis of variance (5, 6, 7) on the sugar beet data.

The original data from table 12.13 in Snedecor's book were not reproduced. The various sums of products from table 12.14 of Snedecor's book are reproduced (table 1) in a slightly different form. The matrix form for presenting the sums of products for the various lines in the analysis of variance is the form commonly used in multivariate analyses (5, 6, 7). The various mean squares are obtained by dividing the sums of products by the appropriate degrees of freedom. The multivariate components in the last column are derived in the usual manner. For example, $77.0015 =$

$$\frac{1}{7} \{ 1494.5140 - 955.5033 \}, \text{ and } 96.1558 = \frac{1}{6} \{ 599.6750 - 22.7400 \}.$$

The multivariate components are useful in much the same manner as are ordinary variance components.

In order to test the significance of a given line in the bivariate analysis of variance it is first necessary to compute a statistic denoted as U. The square root of the quantity U has a beta-distribution and is related to Snedecor's F as follows:

$$F(n, mdf) = \frac{1 - \sqrt{U}}{\sqrt{U}} \left(\frac{p(\text{error df} - 1) = m}{p(\text{treatment df} = n)} \right),$$

where p = number of characteristics. Thus, ordinary variance ratio tables may be used to test the significance of the treatment mean squares, and no new tables are necessary.

The statistic U is computed as described in table 1. For p = 2 characteristics the evaluation of a two 2x2 determinant is all that is required in computing U. The corresponding F is computed as described in the bottom part of table 1, and is compared with the tabulated values of F for 12 and 58 degrees of freedom. If the 5 percent level of significance were being used the calculated F = 7.80 greatly exceeds the tabulated $F_{05} = 1.93$. Hence, there is little doubt that differences in yield and stand for the treatments are present.

In order to obtain some information concerning the nature of the differences we next compute the multivariate (bivariate here) criterion which involves inspection of the various roots. For two characteristics the two roots are obtained as follows:

$$\begin{vmatrix} (E_{11} + V_{11})U - E_{11} & (E_{21} + V_{21})U - E_{21} \\ (E_{12} + V_{12})U - E_{12} & (E_{22} + V_{22})U - E_{22} \end{vmatrix}$$

$$= \begin{vmatrix} 144,685.43U - 28,665.10 & 4,280.25U - 682.20 \\ 4,280.25U - 682.20 & 136.09U - 23.23 \end{vmatrix}$$

$$= 1,369,700.1062U^2 - 1,422,102.8979U + 200,493.4330 = 0, \text{ or}$$

$$U^2 - 1.03825859U + .14637761 = 0.$$

Solution of the above quadratic equation results in the following two roots for U:

$$\begin{aligned} U_1 &= .87001, \\ U_2 &= .16825, \text{ and} \\ U = U_1 U_2 &= .16825(.87001) = .1464. \end{aligned}$$

Table 1. Bivariate analysis of variance on the sugar beet example
(Snedecor, Statistical Methods, p. 332).

Source of variation	df	Sum of products	Mean product	Multivariate components
Total	41	$\begin{pmatrix} 152,158.00 & 4,163.69 \\ 4,163.69 & 142.4022 \end{pmatrix}$		
Replication or block	5	$\begin{pmatrix} 7,472.57 & -116.56 \\ -116.56 & 6.3134 \end{pmatrix}$	$\begin{pmatrix} 1,494.5140 & -23.3120 \\ -23.3120 & 1.2627 \end{pmatrix}$	$\begin{pmatrix} 77.0015 & -6.5789 \\ -6.5789 & 0.0698 \end{pmatrix}$
Treatment	$f_t=6$	$\begin{pmatrix} 116,020.33 & 3,598.05 \\ 3,598.05 & 112.8562 \end{pmatrix}$	$\begin{pmatrix} 19,336.7217 & 599.6750 \\ 599.6750 & 18.8094 \end{pmatrix}$	$\begin{pmatrix} 3,063.5364 & 96.1558 \\ 96.1558 & 3.0058 \end{pmatrix}$
Error	$f_e=30$	$\begin{pmatrix} 28,665.10 & 682.20 \\ 682.20 & 23.2326 \end{pmatrix}$	$\begin{pmatrix} 955.5033 & 22.7400 \\ 22.7400 & 0.7744 \end{pmatrix}$	

$$U = \frac{\begin{vmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{vmatrix}}{\begin{vmatrix} E_{11} + V_{11} & E_{12} + V_{12} \\ E_{21} + V_{21} & E_{22} + V_{22} \end{vmatrix}} = \frac{\begin{vmatrix} 28,665.10 & 682.20 \\ 682.20 & 23.23 \end{vmatrix}}{\begin{vmatrix} 144,685.43 & 4,280.25 \\ 4,280.25 & 136.09 \end{vmatrix}} = \frac{200,493.4330}{1,369,700.1062} = 0.14637761$$

$$F = \frac{1 - \sqrt{U}}{\sqrt{U}} \left(\frac{2(f_e - 1)}{2f_t} \right) = \frac{.6174(58)}{.3826(12)} = \frac{35.8092}{4.5912} = 7.80 \text{ with 12 and 58 degrees of freedom.}$$

Significance of the U's is obtained from the formulae:

$$\begin{aligned} \chi^2(pf_t df) &= - \left\{ f_e + f_t - \frac{1}{2(p + f_t + 1)} \right\} \log_e U \\ &= -2.30259 \left\{ 30 + 6 - \frac{1}{2(2 + 6 + 1)} \right\} \log_{10}(.14637761) = 69.07, \end{aligned}$$

$$\begin{aligned} \chi^2(p + f_t - 1 df) &= - \left\{ f_e + f_t - \frac{1}{2(p + f_t + 1)} \right\} \log_e U_2 \\ &= -2.30259 \left\{ 30 + 6 - \frac{1}{2(2 + 6 + 1)} \right\} \log_{10}(.16825) = 64.06, \end{aligned}$$

and

$$\begin{aligned} \chi^2((p - 1)(f_t - 1) df) &= - \left\{ f_e + f_t - \frac{1}{2(p + f_t + 1)} \right\} \log_e U_1 \\ &= -2.30259 \left\{ 30 + 6 - \frac{1}{2(2 + 6 + 1)} \right\} \log_{10}(.87001) = 5.00, \end{aligned}$$

where f_t = treatment degrees of freedom, f_e = error degrees of freedom, and p = number of characteristics. The above results are summarized in the following table:

Root	df	χ^2	Prob. of greater χ^2
U_1	$(p - 1)(f_t - 1) = 5$	5.00	approx. = .42
U_2	$p + f_t - 1 = 7$	64.06	< .0001
U	$pf_t = 12$	69.07	< .0001

The chi-squares for the two roots, U_1 and U_2 , add to the chi-square value for U , within rounding errors.

Since only one of the roots is significant this means that a single index or criterion may be used to discriminate among the treatment means; i.e., a discriminant function analysis (1, 2, 3) may be used. If both roots were significant this would indicate that a single index could not be used to discriminate among the treatments and that the discriminant function analysis is not sufficient. Instead some multivariate criterion must be used. The fact that a single index may be used to discriminate among the treatments is evident from the close relationship between the treatment means for stand and for yield (see figure 1).

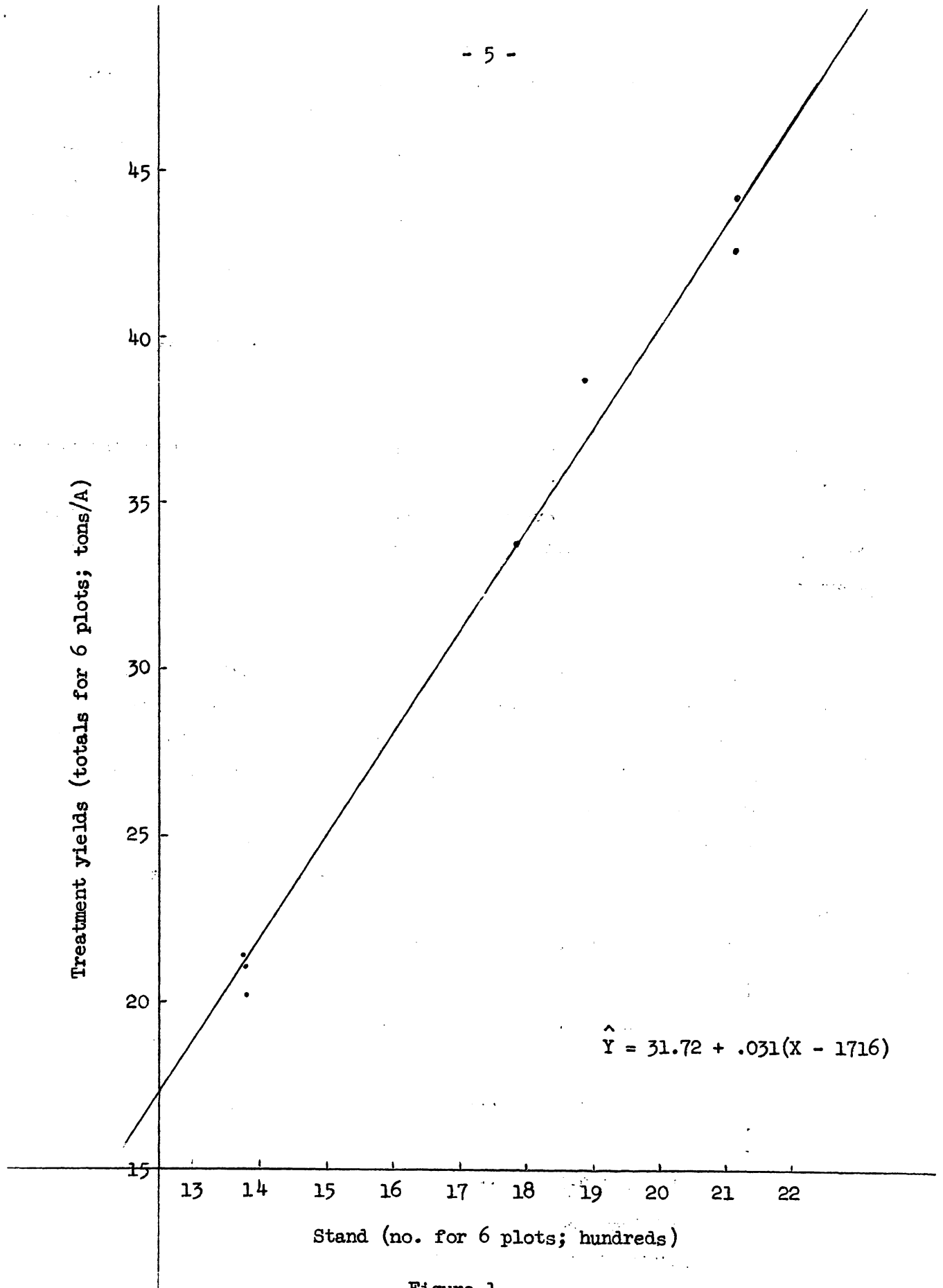


Figure 1.

Smith (3) first described the application of a discriminant function for plant selection. He points out that in selecting for quantitative characters, such as yield, the differences due to genotype are masked by environmental effects. In selection work, plant breeders attempt to select plants on the basis of observable characteristics which they believe are associated with the desired character. The actual worth of each of the observable characteristics is usually unknown. Smith suggested that the discriminant function approach be used to best indicate the "genetic value" of a line.

It may be assumed that the true genotype of the plant is measured by

$$\theta = \sum a_i \xi_i,$$

where the a_i are assigned values representing the relative value of the observed characters X_i whose true values are ξ_i .

This function cannot be evaluated directly because we observe only the phenotypic performance instead of the genotypic performance. Now, let the phenotypic value be represented by the equation

$$Y = \sum b_i X_i,$$

and the problem is to find values of b_i such that the function Y will best discriminate those lines which have the greatest genotypic value θ . That is, the b_i are such that the regression of Y on θ will be a maximum. If the t_{ij} represent the line variances and covariances, the e_{ij} represent the error variances and covariances, and if the $g_{ij} = t_{ij} - e_{ij}$ represent a multiple of the multivariate components which is an estimate of the component due to genotype, maximization of the regression of Y on θ results in the following equations:

$$\begin{aligned} b_1 t_{11} + b_2 t_{12} + \dots + b_p t_{1p} &= A_1 \\ \vdots & \\ b_1 t_{1p} + b_2 t_{2p} + \dots + b_p t_{pp} &= A_p, \end{aligned}$$

where

$$\begin{aligned} A_1 &= a_1 g_{11} + a_2 g_{12} + \dots + a_p g_{1p} \\ \vdots & \\ A_p &= a_1 g_{1p} + a_2 g_{2p} + \dots + a_p g_{pp}. \end{aligned}$$

The A_j are computed from the data after the a_i have been decided upon.

Goulden (2) suggests that the a_i be set equal to the reciprocal of some

multiple of the mean for a given character. At first sight it would appear that such a procedure would only be valid for characters measured on the same scale, and if different units are used the procedure is not invariant.

The above procedure may be used directly on the sugar beet data after the relative worth of the two characters, stand and yield per plot, are decided upon. Since stand is such an important part of yield in sugar beets it might be advisable to give them equal weights, that is, $a_1 = 1 = a_2$. Any other values that appear reasonable to the experimenter may be used. From table 1 the $p = 2$ equations involving b_1 and b_2 are:

$$19,336.7217 b_1 + 599.6750 b_2 = 3,063.5364 a_1 + 96.1558 a_2 = 3,159.6922.$$

$$599.6750 b_1 + 18.8094 b_2 = 96.1558 a_1 + 3.0058 a_2 = 99.1616.$$

Solution of the above two equations for b_1 and b_2 results in the following:

$$b_1 = - .008000436 \quad \text{and}$$

$$b_2 = 5.526984463 ,$$

or in relative values

$$b_1 \sqrt{E_{11}/f_e} = - .008000436 \sqrt{955.5033} = .2473 \quad \text{and}$$

$$b_2 \sqrt{E_{22}/f_e} = 5.526984463 \sqrt{.7744} = 4.8637.$$

Thus, yield is about 20 times more important than stand in discriminating among the 7 treatments. This is not unexpected since yield per plot is equal to stand times average weight per beet. The discriminant function is

$$Y = - X_1 + 690.8 X_2 ,$$

where the second regression coefficient is in units of the first (lowest) coefficient.

The bivariate analysis of average weight per beet* and stand per plot (table 2) is given in table 3. The corresponding U value and the two roots U_1 and U_2 are given in table 3. Since only one of the roots is significant a single discriminant function may be used (see figure 2). The two equations involving b_1 and b_2 for these data and for $a_1 = 1 = a_2$ are:

$$19,336.7 b_1 + .9156 b_2 = 3,063.53 a_1 + .1486 a_2 = 3,063.6786 \quad \text{and}$$

$$.9156 b_1 + .00004717 b_2 = .1486 a_1 + .00000719 a_2 = .14860719 ,$$

*This is the average weight per beet multiplied by a factor to convert to tons per acre. The multiplication by a constant does not change the relative values of the characters.

Table 2. Number of beets per plot and average yield per plot
for Snedecor's sugar beet example.

Fertilizer applied	No. and avg. yield	Blocks						Totals
		1	2	3	4	5	6	
None	No. = X_1	183	176	291	254	225	249	1378
	Avg. yield = X_2	.0134	.0128	.0151	.0171	.0152	.0131	.0867
P = super- phosphate	X_1	356	300	301	271	288	258	1774
	X_2	.0188	.0181	.0163	.0193	.0234	.0184	.1143
K = muriate of potash	X_1	224	258	244	217	192	236	1371
	X_2	.0144	.0160	.0095	.0204	.0171	.0169	.0943
P + K	X_1	329	283	308	326	318	318	1882
	X_2	.0193	.0192	.0169	.0245	.0219	.0219	.1237
P + N (N = sodium nitrate)	X_1	371	354	352	331	290	410	2108
	X_2	.0175	.0201	.0167	.0228	.0228	.0216	.1215
K + N	X_1	230	221	237	193	247	250	1378
	X_2	.0161	.0147	.0119	.0111	.0210	.0165	.0913
P + K + N	X_1	322	367	400	333	314	385	2121
	X_2	.0189	.0209	.0184	.0235	.0247	.0192	.1256
Totals	X_1	2015	1959	2133	1925	1874	2106	12012
	X_2	.1184	.1218	.1048	.1387	.1461	.1276	.7574

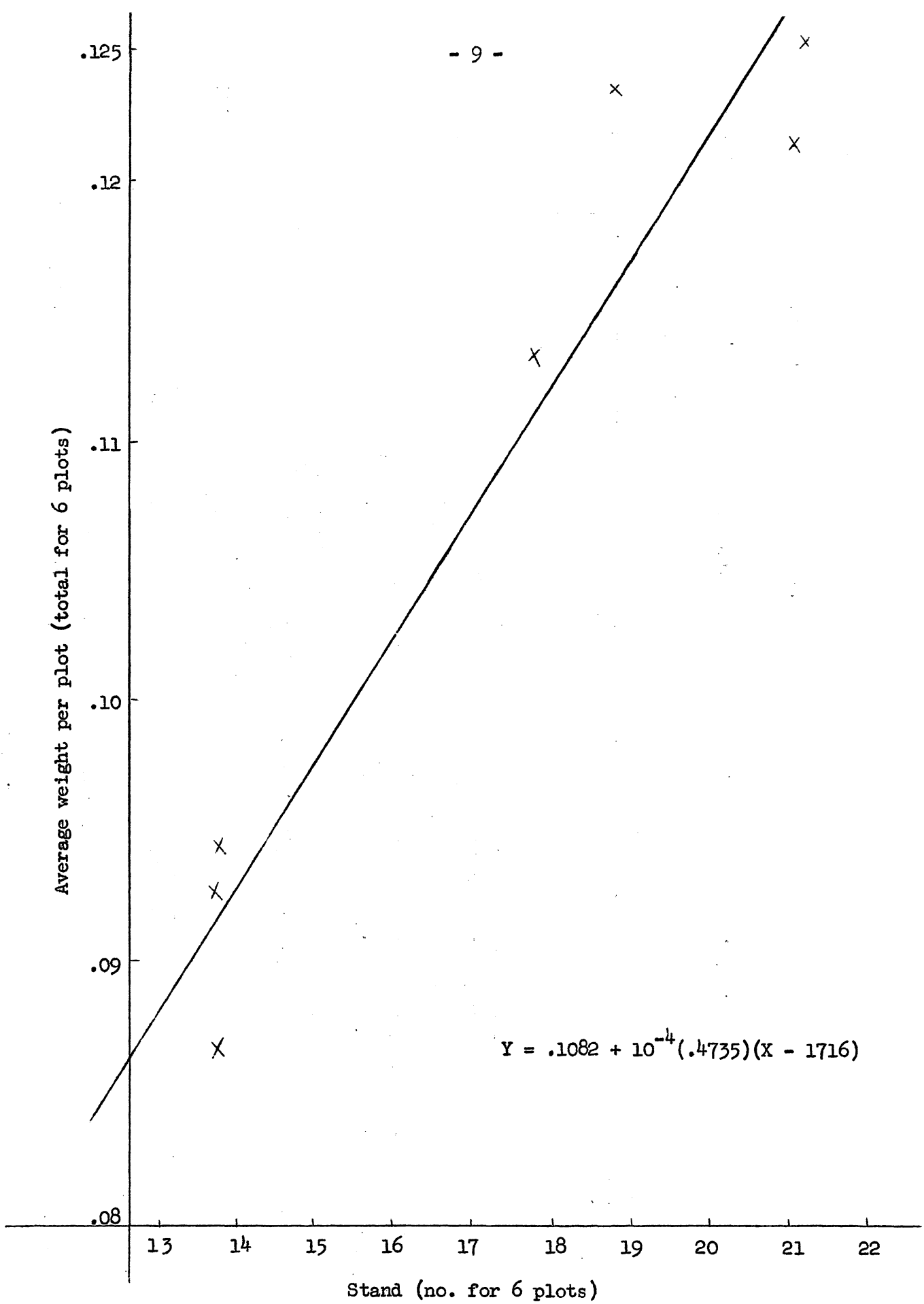


Figure 2.

Table 3. Bivariate analysis of variance for mean weight per sugar beet per plot and stand per plot for the sugar beet example (Snedecor, Statistical Methods, p. 332).

Source of variation	df	Sum of products	Mean product	Multivariate Components
Total	41	$\begin{pmatrix} 152,158 & 5.3445 \\ 5.3445 & 0.00055969 \end{pmatrix}$		
Replicate or block	5	$\begin{pmatrix} 7,473 & -0.8685 \\ -0.8685 & 0.00015605 \end{pmatrix}$	$\begin{pmatrix} 1,494.6 & -0.1737 \\ -0.1737 & 0.00003121 \end{pmatrix}$	$\begin{pmatrix} 77.01 & -0.0282 \\ -0.0282 & 0.00000388 \end{pmatrix}$
Treatment	$f_t = 6$	$\begin{pmatrix} 116,020 & 5.4937 \\ 5.4937 & 0.00028303 \end{pmatrix}$	$\begin{pmatrix} 19,336.7 & 0.9156 \\ 0.9156 & 0.00004717 \end{pmatrix}$	$\begin{pmatrix} 3,063.53 & 0.1486 \\ 0.1486 & 0.00000719 \end{pmatrix}$
Error	$f_e = 30$	$\begin{pmatrix} 28,665 & 0.7193 \\ 0.7193 & 0.00012061 \end{pmatrix}$	$\begin{pmatrix} 955.5 & 0.0240 \\ 0.0240 & 0.00000402 \end{pmatrix}$	

$$U = \frac{\begin{vmatrix} 28665 & 0.7193 \\ 0.7193 & 0.00012061 \end{vmatrix}}{\begin{vmatrix} 144685 & 6.2130 \\ 6.2130 & 0.00040364 \end{vmatrix}} = \frac{2.93989316}{19.79928440} = 0.14848482 ; \quad F(12, 58 \text{ df}) = \frac{58}{12} \left(\frac{1 - .38533}{.38533} \right) = 7.71.$$

$$19.79928440 U^2 - 20.08277665 U + 2.93989316 = 0; \quad U_2 = .1774; \quad U_1 = .8369.$$

$$\chi^2_{U_2} (7 \text{ df}) = 62.16; \quad \chi^2_{U_1} (5 \text{ df}) = 6.40.$$

and the solution for the b's is:

$$\begin{aligned}b_1 &= .114514341 \quad \text{and} \\b_2 &= 927.629246 ,\end{aligned}$$

or in relative values

$$\begin{aligned}b_1 \sqrt{E_{11}/f_e} &= .114514341 \sqrt{955.5} = 3.5398 \quad \text{and} \\b_2 \sqrt{E_{22}/f_e} &= 927.629246 \sqrt{.00000402} = 1.8599 .\end{aligned}$$

The discriminant function is

$$Y = X_1 + 8101 X_2 .$$

In the above analysis stand per plot is about twice as important as average weight per beet in discriminating among the 7 treatments. As a further analysis one could use the two characters $\log X_1$ and $\log X_2$. This may be more realistic than using X_1 and X_2 because $\text{yield} = X_1 X_2$, whereas $\log \text{yield} = \log X_1 + \log X_2$. The analysis on logarithms is left as an exercise for the reader.

In both figures 1 and 2 it should be noted that the fertilizer treatments containing phosphorus had a considerable effect on both stand and average weight per beet.

References

1. Brown, G., Discriminant functions. Ann. Math. Stat. 18:514-528, 1947.
2. Goulden, C. H., Methods of statistical analysis, chapter 17, 2nd edition. Wiley, N. Y., 1952.
3. Smith, H. F., A discriminant function for plant selection. Ann. Eugenics 7:240-250, 1936-7.
4. Snedecor, G. W., Statistical methods, section 12.7, 4th edition. Iowa State College Press, Ames, 1946.
5. Steel, R. G. D., On multivariate analysis. Mimeo BU-50-M, Cornell Univ., June, 1952.
6. Steel, R. G. D., Multivariate analysis of yield in perennial experiments. Paper presented at Biometric Society, ENAR, Meeting in Madison, Wisconsin, Sept. 7, 1953.
7. Tukey, J. W., Dyadic anova, an analysis of variance for vectors. Human Biology 21:65-110, 1949.